GANtruth – an unpaired image-to-image translation method for driving scenarios

Sebastian Bujwid
KTH Royal Institute of Technology
Univrses AB
bujwid@kth.se

Miquel Martí
KTH Royal Institute of Technology
Univrses AB
miquelmr@kth.se

Hossein Azizpour
KTH Royal Institute of Technology
azizpour@kth.se

Alessandro Pieropan
Univrses AB
alessandro.pieropan@univrses.com

Abstract

Synthetic image translation has significant potentials in autonomous transportation systems. That is due to the expense of data collection and annotation as well as the unmanageable diversity of real-words situations. The main issue with unpaired image-to-image translation is the ill-posed nature of the problem. In this work, we propose a novel method for constraining the output space of unpaired image-to-image translation. We make the assumption that the environment of the source domain is known (e.g. synthetically generated), and we propose to explicitly enforce preservation of the ground-truth labels on the translated images.

We experiment on preserving ground-truth information such as semantic segmentation, disparity, and instance segmentation. We show significant evidence that our method achieves improved performance over the state-of-the-art model of UNIT for translating images from SYNTHIA to Cityscapes. The generated images are perceived as more realistic in human surveys and outperforms UNIT when used in a domain adaptation scenario for semantic segmentation.

1 Introduction

Autonomous driving has a high potential for improving human life in many aspects, such as reducing the number of accidents and giving commuters the choice to dedicate their travel time to other activities. However, due to the diverse nature of the environment and the unpredictable behavior of humans we are still far from solving the problem and a particular effort is needed in order to enable a reliable understanding of the surroundings, essential for decision making. In this regard, deep learning techniques have shown impressive results in many tasks such as object detection, tracking and semantic segmentation. These methods, though, require vast amounts of data, which is very expensive to collect and annotate. An appealing solution lies in using synthetic data instead [19, 12, 18], given that the ground-truth annotation can be extracted automatically from the simulators. Such an approach has shown promising results in applications like human pose estimation and hand tracking [4, 22].

Synthetic data has enormous potential especially in autonomous driving, because safety critical situations that are exceptionally rare in real life can be simply simulated. Such flexibility also allows to render data corresponding to various environments and conditions which make it easier to adapt cars to different countries where not only the environment, but also regulations might be different. Given the added value of synthetic data, our work aims to improve it's realism by finding a mapping between real and synthetic domains using generative models to translate unpaired sets of images.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

However, since the problem is generally ill-posed and thus can have infinitely many solutions, in this work we aim to constrain the output space by the preservation of the content that should be representation invariant in the special case of synthetic-to-realistic translation and thereby improve the translation utility. We enforce the preservation of synthetic domain ground-truth information in the translated images by means of a loss between the source domain ground-truth labels and the output of a pre-trained label estimator on the target domain. The results of the method we propose are judged as more realistic by human evaluators and achieve better performance on a standard domain adaptation scenario than the state-of-the-art model of UNIT.

2 Related work

Image-to-image translation can be considered as a variation of the visual adaptation problem firstly introduced by Saenko et al. [21]. In recent years, using generative adversarial networks (GANs) [7] to synthesize images has shown promising results. The main idea lies in training two competing networks, while the generator learns to produce realistic images, the discriminator network learns to determine if an image is real or not. However, in GANs the output depends on the input random noise vector leaving no semantic control on the generated data. A proposed solution to the former is to condition the output on the source domain samples leading to many interesting applications in image-to-image translation such as the generation of natural scenes [26], super resolution of images [25] or image manipulation [30, 11]. Some methods proposed in recent years build upon the assumption that images in the source domain have correspondences with the images in the target domain. With this assumption, Isola et al. [10] proposed a network architecture and loss function and produced impressive results. Wang et al. [25] has shown that it is possible to generate high resolution images from semantic segmentation maps by learning the mapping between the domains. Despite the impressive results of the mentioned methods, the assumption of having correspondences between the domains is quite limiting due to the scarcity of manually labeled data or the nature of the chosen domain pairs themselves, e.g. depth maps with natural images or synthetic images with natural images. In order to overcome this limitation Liu et al. [17] proposed forcing a shared latent space between the domains in the framework called UNIT. Moreover, they enforced cycle consistency as introduced in CycleGAN [31], DiscoGAN [13] and DualGAN [27], so that a generated image in the target domain could be mapped back to the source domain.

In this work, we are interested in preserving the domain-agnostic content of the images. Previous works in this direction relied on jointly learning the task network on the translated images [29, 9]. The problem with such an approach is that it does not necessarily have to force the translated images to be consistent with the source samples if only the task networks learn to ignore the content that was modified. CyCADA [9] is conceptually closest to our work in that it also tries to preserve semantic consistency during the translation. CyCADA proposes to minimize the discrepancy between labels estimated by a model trained on source domain. We argue that this approach can be problematic since it assumes the label estimators trained on the source domain is re-usable for the target domain. This assumption essentially violates the definition of the domain adaptation (*i.e.* shift in the visual representation).

3 GANtruth

Assume an image space $\mathbb{X}$ and two probability distributions $P_S(x)$ and $P_T(x)$ over $\mathbb{X}$ denoting the source and target image domains respectively. In unpaired image-to-image translation, given the training sets $X_S = \{x_S \sim P_S(x)\}$ and $X_T = \{x_T \sim P_T(x)\}$, we are interested in learning an adaptation function $F_{S\to T}$ such that $F_{S\to T}(x \sim P_S) \sim P_T$ [1]

In general, this is an ill-posed problem where several valid solutions can exist. Also, in the particular case of synthetic-to-natural image translation many (potentially infinite) valid natural images might exist for a given synthetic image . Motivated by this problem, we propose to constrain the learning problem by requiring the image content to be preserved during the translation. Formally, assume a label space $\mathbb{Y}$, we require a translation function that additionally satisfies the following constraint.

$$P(y|x_S) = p(y|F_{S\to T}(x_S)) \quad \forall x_S \sim P_S \tag{1}$$

[1] In its general form, the translation function can produce $P(x_T|x_S)$ instead of a deterministic mapping.

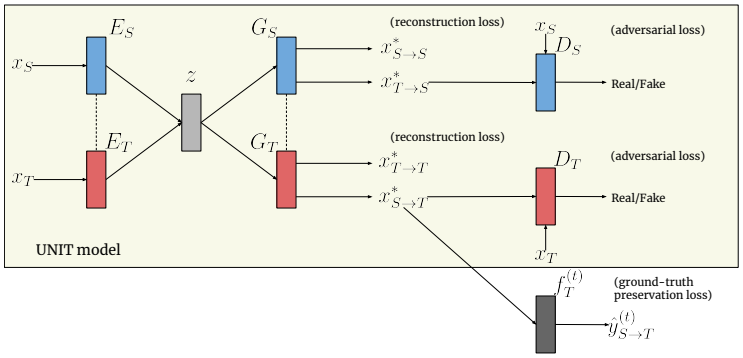


Figure 1: **Diagram of GANtruth based on UNIT framework (GANtruth+UNIT)**. The model takes as an input a sample from each domain and encodes the samples into the same latent space. Then, from the latent space, those samples can be translated back to the same domain using the decoder of the same domain as the original sample, or translated to the other domain using the other decoder. The blue modules work in the source domain, while the red ones in the target domain and dashed lines mean that weights are partially shared.

This formulation assumes access to the label-conditional in both domains. We relax the constraint in the following *ground-truth preservation loss* function:

$$\mathcal{L}_{GT} = \sum_{x_S \in X_S} \mathcal{L}(f_T(F_{S\to T}(x_S)), y_S) \tag{2}$$

where $\mathcal{L}$ is some distance measure over $\mathbb{Y}$. The loss function assumes labels for the source images (y_S). Our source images are synthetically generated which come with free information about the environment (used to render x_S). We also assume *some* pre-trained estimator (f_T) for the target domain exists. In practice, some tasks are easier to solve than others, so we can leverage them for solving more difficult ones. Several ground-truth preservation losses can be aggregated if different pre-trained estimators ($f_T^{(t)}$) are available for the target domain. In fact, this is usually the case in synthetic-to-real images, since several off-the-shelf visual recognition (*e.g.* object detection, semantic segmentation, depth estimation, instance segmentation) models are available for natural images.

3.1 Translation Models

We examine the effectiveness of the ground-truth preservation loss ($\mathcal{L}_{GT}$) on a basic GAN-based [7] model, minimizing the following objective:

$$\mathcal{L}_{GANtruth} = \lambda_{GAN} \cdot \mathcal{L}_{GAN}(E_S, G_T, D_T) + \mathcal{L}_{GT}(f_T, F_{S\to T}, x_S, y_S). \tag{3}$$

Additionally, we also study the impact of the ground-truth preservation loss on a state-of-the-art image-to-image translation model called UNIT [17]. UNIT is a bidirectional translation model which works through a shared latent space. The UNIT's translation functions consist of an encoder and decoder for each domain denoted by E_T, E_S and G_T, G_S respectively and is defined as $F_{S\to T}(x_S) = G_S(z \sim \mathcal{N}(E_S(x_S), I))$. UNIT's objective function augmented with the ground-truth preservation loss becomes:

$$\begin{aligned}\mathcal{L}_{GANtruth+UNIT} &= \mathcal{L}_{VAE}(E_S, G_S) + \lambda_{GAN} \cdot \mathcal{L}_{GAN}(E_S, G_S, D_S) + \mathcal{L}_{CC}(E_S, G_S, E_T, G_T) \\ &+ \mathcal{L}_{VAE}(E_T, G_T) + \lambda_{GAN} \cdot \mathcal{L}_{GAN}(E_T, G_T, D_T) + \mathcal{L}_{CC}(E_T, G_T, E_S, G_S) \\ &+ \mathcal{L}_{GT}(f_T, F_{S\to T}, x_S, y_S),\end{aligned} \tag{4}$$

with D_T, D_S being discriminator functions that classify images into the two domains. The UNIT objective has several loss terms: Varitional AutoEncoder (VAE) [14] losses denoted by $\mathcal{L}_{VAE}$ to ensure reconstruction from the latent space for source and target images, Generative Adversarial Network (GAN) [7] losses denoted by $\mathcal{L}_{GAN}$ which try to make the translated source images indistinguishable from target images and vice versa, and cycle consistency [31] losses (L_{CC}) which additionally encourage perfect reconstruction of images after a translation to the other domain and back. In Figure 1 we show a diagram of the augmented model. The parameters of f_T are not updated during training. The details of the losses can be found in Appendix A.

4 Experiments

We experiment with our method using 16212 stereo images of *Summer* and *Spring* sequences from SYNTHIA-Sequences dataset [20] as a source domain. As a target domain we use Cityscapes [3] train and train-extra sets, containing 23472 images. The setup is similar to the one used by Liu et al. [17]. The details of the network architecture are described in Appendix E.

4.1 Enabling preservation of different ground-truth

We experiment with enforcing preservation of different types of labels. Different modules corresponding to different ground-truth information are enabled in order to observe their individual impact, as well as the results of combining them together.

Semantic segmentation For preserving semantic information we use the semantic segmentation network ICNet [28] as f_T and cross entropy loss between ground-truth source labels and the estimated labels on $F_{S\to T}(x_S)$. In most of our experiments the model is pre-trained on Cityscapes training and validation set, except experiments in Section 4.4 where only training set was used to make the evaluation fair. Because of the differences between classes that are defined in our source and target datasets, we use the mapping defined in Appendix D.

Depth (disparity) For preserving depth information we use the monocular depth estimation network proposed by Godard et al. [6]. This is an unsupervised model that uses stereo frames during training and produces disparities from both images, which are used to reconstruct the corresponding left or right image. The model is trained to minimize the error of image reconstruction, as well as to preserve consistency between disparities produced from the left and right image. Finally, during test time, the model estimates disparities from just a single (mono) image. The variant of the model we use is based on the ResNet-50 architecture and is trained on the Kitti dataset [5], which we believe is similar enough to our target domain dataset. Predicted disparities are multiplied by a constant to compensate for different camera parameters. The discrepancy measure we use for depth preservation is mean absolute error.

Instance segmentation In order to preserve better fine details of specific objects relevant to autonomous driving scenarios, we enforce preservation of ground-truth labels in the instance segmentation task, which gives a different mask for each individual object in the scene. We choose a Mask-RCNN [8] instance segmentation network based on Inception-v2 [24] and use the losses defined in [8]. Instead of using a model trained on Cityscapes, we use one trained on Microsoft-COCO [15], whose images should belong to a domain much closer to the target domain than the synthetic images of the source domain. See the mapping between classes across datasets in Appendix D.

Hyperparameters We use Adam optimizer with learning rate equal 0.0001 and $\beta_1 = 0.5$, $\beta_2 = 0.999$ and batch size equal 1. We weight the loss function using the following coefficients: $\lambda_{sem.seg.} = 40.0$, $\lambda_{disp.} = 0.4$, $\lambda_{inst.seg.} = 1.0$, $\lambda_{GAN} = 10.0$.

4.2 Human surveys on perceptual realism

To measure perceptual quality of the translated images we use human judges. We run human surveys in the form of A/B tests, using the Amazon Mechanical Turk platform. The participants are presented with two images and are asked to choose which one looks more realistic. The images we show are outputs of different models from synthetic to real translation. This way we measure a relative improvement of image quality against a baseline model.

The input images chosen for the survey are sampled randomly from the dataset and are the same for all the studies. The order of the images (left vs. right) is selected randomly for each comparison.

The results in Table 1 show that images translated by GANtruth with both semantic segmentation and depth preservation are perceived as more realistic than UNIT on average, for images from both datasets. In the case of using only one of the modules available results are only comparable to UNIT in the case of semantic segmentation, but only on the SYNTHIA-Seq dataset. This can indicate that the different modules working at the same time help improving the generalization of the model.

Table 1: AMT survey results, 30 images randomly sampled from the dataset are compared by 15 participants (450 responses) each. Numbers are average values. S = Semantic Segmentation, D = Depth, I = Instance Segmentation, S-SEQ = SYNTHIA-Seq, S-RAND = SYNTHIA-RAND-CVPR16

	Preference against UNIT [17]	
Method	S-SEQ	S-RAND
Simple GAN (baseline)	25.11%	12.22%
GANtruth (S)	61.33%	23.56%
GANtruth (D)	30.22%	8.44%
GANtruth (I)	38.44%	23.56%
GANtruth (S+D)	**61.78%**	**59.11%**
UNIT + GANtruth (S)	64.00%	**78.89%**
UNIT + GANtruth (D)	57.55%	43.33%
UNIT + GANtruth (S+D)	**70.22%**	70.00%
UNIT + GANtruth (S+D+I)	65.78%	61.11%

Table 2: Adaptation for semantic segmentation. Models are evaluated on Cityscapes validation set. S = Semantic Segmentation, D = Depth

Dataset	mIOU (%)
Source domain (SYNTHIA-Seq)	24.9
Simple GAN (baseline)	20.3
UNIT (baseline)	23.0
GANtruth (S+D)	**26.6**
Target domain (Cityscapes-train set)	57.0

Furthermore, GANtruth improves humans' preference on all but one case when combining it with UNIT over only using UNIT, showing that the approaches are complementary.

4.3 Qualitative results

In Figure 2 we show qualitative results for the different single-task GANtruth experiments, the combination of all considered tasks, and the later together with UNIT. In addition, we provide results for two baselines - a simple GAN and UNIT - for comparison. The superior performance of GANtruth with semantic segmentation labels preservation is confirmed in the qualitative results. Results of this method appear free of the largest artifacts in the baseline methods: objects melting into others in simple GAN, black sky in UNIT and simple GAN, vegetation on buildings in UNIT. However, colours of the source domain are not kept. The combination of GANtruth with UNIT keeps source image colours while removing the largest artifacts in UNIT. More images are shown in Appendix B.

4.4 Adaptation for semantic segmentation

We evaluate our approach in a simple, two-step adaptation scenario. First, we construct new datasets made of synthetic data translated using different $F_{S \to T}$ models and corresponding ground-truth labels y. Then, we use these datasets to train a supervised task network for the semantic segmentation problem. Unlike other approaches [9, 29], our adaptation scenario uses two-separate steps – we do not learn the task network during adaptation.

The task model we use to train on adapted images is DeepLabv3+ [1] with Xception [2] backbone and output stride of 32. We fine-tune the model previously pre-trained on ImageNet for 90000 updates on 449x449 image crops with batch size 16. Batch normalization parameters are not updated. We find that these values give good results on the target domain dataset and we use it for all experiments. In Table 2 we present mIOU (mean Intersection-over-Union) scores on unseen Cityscapes validation set and show that training on translated images improves performance over training on synthetic data.

4.5 Additional observations

During our experimentation we noticed that applying image-to-image translation methods does not guarantee improvement on a simple, two-step domain adaptation. This is the case for both baseline models, as well as some variants of our model. Even though it is likely that the results would differ if a more advanced adaptation method was used (e.g. based on aligning feature distributions of the task network) instead of simply using the translated images, we believe that this is an interesting observation. It suggests that the translation models despite producing visually appealing images might not necessarily be directly useful for training machine learning models as it was shown by

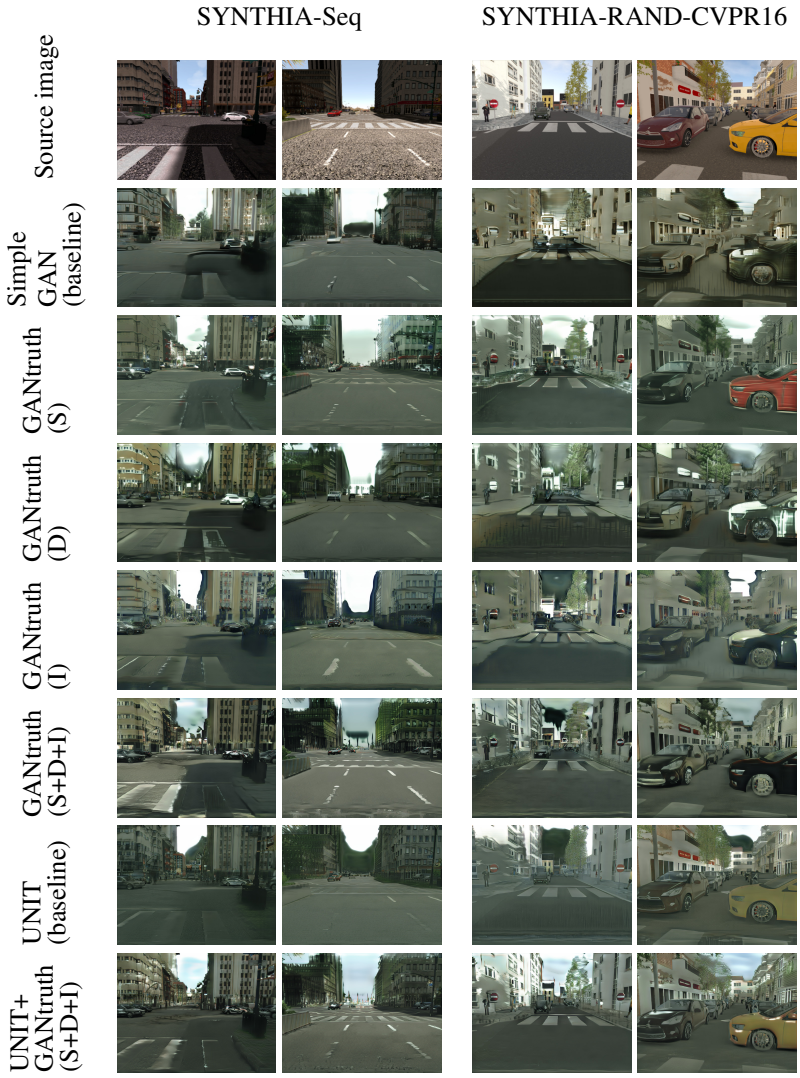


Figure 2: Qualitative results. Images randomly sampled. S = Semantic Segmentation, D = Depth, I = Instance Segmentation.

Shrivastava et al. [23] on datasets that have much smaller domain gaps. Additionally, we observe that when our semantic preservation loss is used, adding semantic consistency loss as proposed in CyCADA [9] degrades the results. This behaviour is somehow expected, since their approach to enforce semantic consistency uses a source domain estimator which cannot be expected to work well on images resembling the target domain. For more details, please refer to Appendix C.

5 Conclusions

In this work, we propose a novel approach for unpaired synthetic-to-realistic translation method on driving scenarios. Our method based on explicitly enforcing preservation of ground-truth information is shown to be an effective approach for constraining the output space of the ill-posed problem of unpaired image-to-image translation. Quantitative results from human surveys indicate that our model using both semantic segmentation and depth information preservation is perceived as more realistic than UNIT, but also that the two approaches are complementary and when combined produce even better results. Additionally, we show that our method can be used for domain adaptation and using translated images for training semantic segmentation network improves the performance over using the original synthetic images.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[2] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[4] Oscar M. Danielsson and Omid Aghazadeh. Human pose estimation from rgb input using synthetic training data. *CoRR*, abs/1405.1213, 2014.

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle consistent adversarial domain adaptation. In *International Conference on Machine Learning (ICML)*, 2018.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

[12] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? In *International Conference on Robotics and Automation*, pages 746–753, 05 2017.

[13] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1857–1865, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[14] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *The 2nd International Conference on Learning Representations (ICLR)*, 2013.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014. URL `https://www.microsoft.com/en-us/research/publication/microsoft-coco-common-objects-in-context/`.

[16] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.

[17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems 30*, pages 700–708. 2017.

[18] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[19] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.

[20] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[21] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV, 2010.

[22] Jamie Shotton, Andrew Fitzgibbon, Andrew Blake, Alex Kipman, Mark Finocchio, Bob Moore, and Toby Sharp. Real-time human pose recognition in parts from a single depth image. IEEE, June 2011. URL `https://www.microsoft.com/en-us/research/publication/real-time-human-pose-recognition-in-parts-from-a-single-depth-image/`.

[23] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 6, 2017.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[25] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[26] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[27] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. *arXiv preprint arXiv:1704.08545*, 2017.

[29] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2223–2232, 2017.

A UNIT

Instead of explicitly defining the relationship between images in different domains, UNsupervised Image-to-image Translation (UNIT) [17] proposed to create a shared, domain-agnostic latent space. Based on a previously introduced Coupled GAN (CoGAN) [16], they proposed an architecture consisting of two Variational Auto-encoders (VAEs) [14] and two GANs. The model simultaneously learns to translate in two directions. Each encoder part of the VAEs corresponds to a different domain and both of them encode images into the same, shared latent space (z–space). Additionally, decoder parts of each VAEs are appended with separate discriminators for given domains. Combining those parts gives a separate GAN for each domain.

The framework allows to transform images in different directions, e.g. an image in domain A is encoded into a z–space and after that can be decoded either into domain B or back again into domain A. The whole model is jointly trained to reconstruct the images with VAEs, translate images with GANs and additionally preserve cycle-consistency.

The UNIT's translation functions consist of an encoder and decoder for each domain denoted by E_T, E_S and G_T, G_S respectively and is defined as,

$$\begin{aligned} x^*_{S\to T} &= F_{S\to T}(x_S) = G_T(z \sim \mathcal{N}(E_S(x_S), I)) \\ x^*_{T\to S} &= F_{T\to S}(x_T) = G_S(z \sim \mathcal{N}(E_T(x_T), I)). \end{aligned} \tag{5}$$

The objective function of UNIT consists of various terms and is given by:

$$\begin{aligned} \mathcal{L}_{UNIT} =& \mathcal{L}_{VAE}(E_S, G_S) + \lambda_{GAN} \cdot \mathcal{L}_{GAN}(E_S, G_S, D_S) \\ &+ \mathcal{L}_{CC}(E_S, G_S, E_T, G_T) \\ &+ \mathcal{L}_{VAE}(E_T, G_T) + \lambda_{GAN} \cdot \mathcal{L}_{GAN}(E_T, G_T, D_T) \\ &+ \mathcal{L}_{CC}(E_T, G_T, E_S, G_S) \end{aligned} \tag{6}$$

with D_T, D_S being discriminator functions that classify images into the two domains. Generative Adversarial Network (GAN) [7] losses denoted by $\mathcal{L}_{GAN}$ try to make the translated source images indistinguishable from target images and vice versa. The GAN losses are defined as:

$$\begin{aligned} \mathcal{L}_{GAN} &= \mathcal{L}^{(D)}_{GAN} + \mathcal{L}^{(G)}_{GAN} \\ \mathcal{L}^{(D)}_{GAN} &= -\frac{1}{2}\,\mathbb{E}_{\mathbf{x}\sim p_{data}} \log D(\mathbf{x}) - \frac{1}{2}\,\mathbb{E}_{\mathbf{z}} \log(1 - D(G(\mathbf{z}))) \\ \mathcal{L}^{(G)}_{GAN} &= -\frac{1}{2}\,\mathbb{E}_{\mathbf{z}} \log(D(G(\mathbf{z}))). \end{aligned} \tag{7}$$

Varitional AutoEncoder (VAE) [14] losses denoted by $\mathcal{L}_{VAE}$ ensure reconstruction from the latent space for source and target images and is formally defined by

$$\begin{aligned} \mathcal{L}_{VAE}(E, G) =& \lambda_{kl} \cdot D_{KL}(\mathcal{N}(E(x), I) \| \, \mathcal{N}(0, I)) \\ &- \lambda_{ll}\,\mathbb{E}_{z\sim\mathcal{N}(E(x),I)}[\log p_G(x|z)] \end{aligned} \tag{8}$$

Cycle consistency [31] losses (L_{CC}) additionally encourage perfect reconstruction of images after a translation to the other domain and back. The cycle-consistency loss is defined as:

$$\begin{aligned} \mathcal{L}_{CC}(E_S, G_S, E_T, G_T) =& \lambda_{kl} \cdot D_{KL}(\mathcal{N}(E_S(x_S), I) \| \, \mathcal{N}(0, I)) \\ &+ \lambda_{kl}(\mathcal{N}(E_T(G_T(z \sim \mathcal{N}(E_S(x_s), I)), I) \| \, \mathcal{N}(0, I)) \\ &- \lambda_{ll}\,\mathbb{E}_{z\sim\mathcal{N}(E_T(G_T(z\sim\mathcal{N}(E_S(x_S),I)),I)}[\log p_{G_T}(x_S|z)] \end{aligned} \tag{9}$$

Finally, it is worth mentioning, that the weights of the last layers of the encoders and the first layers of the decoders are shared between domains. This helps to preserve a universal semantic meaning of the latent space.

B Additional images

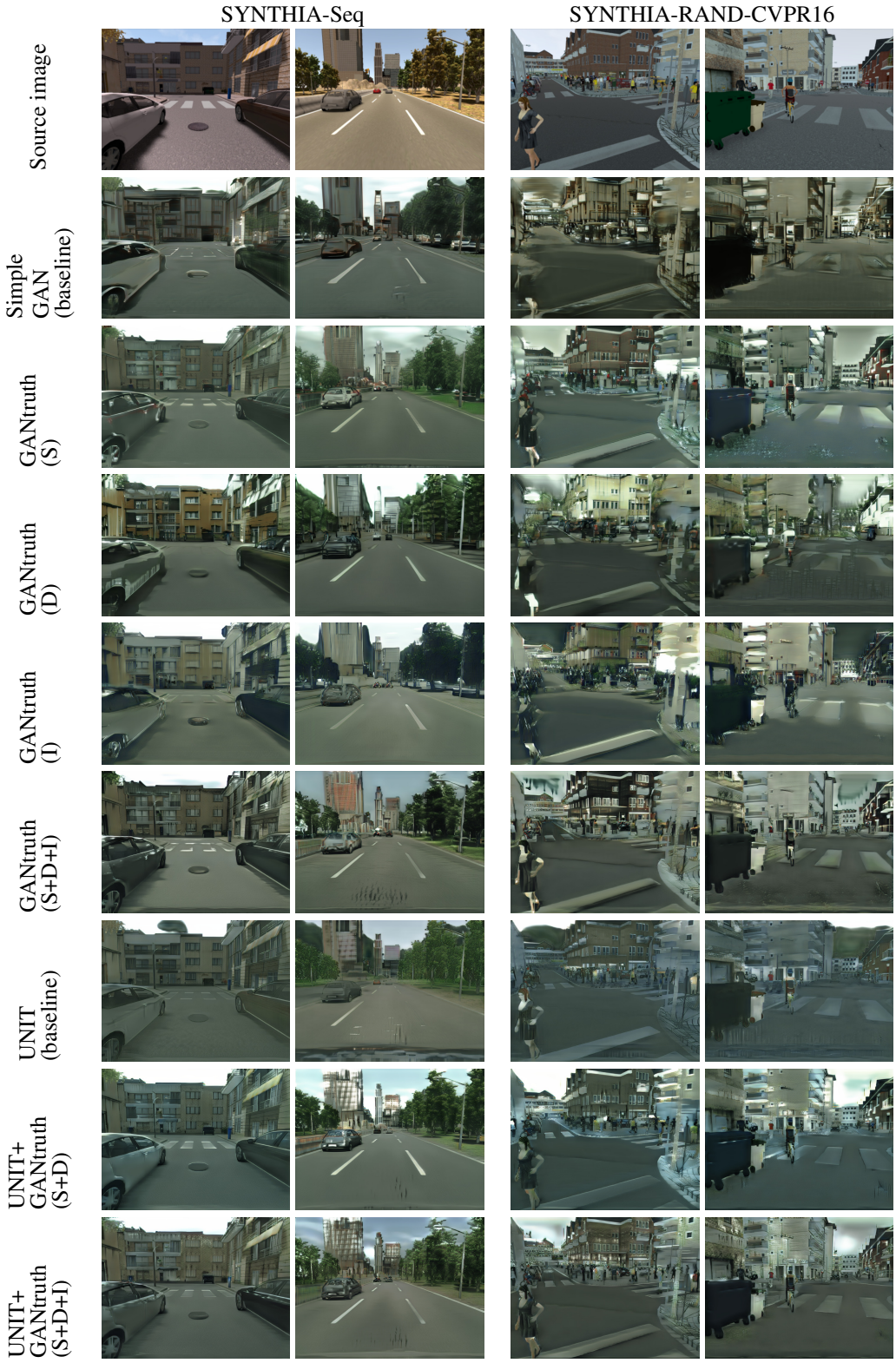


Figure 3: Qualitative results. Images randomly sampled. S = Semantic Segmentation, D = Depth, I = Instance Segmentation.

Table 3: Adaptation for semantic segmentation. S = Semantic Segmentation, D = Depth, I = Instance Segmentation.

Dataset	mIOU (%)
Source domain (SYNTHIA-Seq)	24.9
Simple GAN (baseline)	20.3
GANtruth (S+D)	**26.6**
GANtruth (D+I)	21.5
GANtruth (S+D) + sem.consistency [9]	24.4
GANtruth (D+I) + sem.consistency [9]	25.2
UNIT (baseline)	23.0
UNIT + GANtruth (D+I)	22.9
Target domain (Cityscapes)	**57.0**
Target domain + GANtruth (S+D) + sem.consistency	40.3
Target domain + GANtruth (D+I) + sem.consistency	40.6

C Additional adaptation results

C.1 Adaptation for semantic segementation

More results from experiments described in Section 4.4 are shown in Table 3. A possible explanation of why our observations differ from results presented in SimGAN [23] is that in driving datasets, the gap between domains is much higher. This is evident when comparing the accuracies of models trained on source versus target domains. The difference between them was only marginal on the datasets used by Shrivastava et al. [23], while at the same time the amount of synthetic data was greatly exceeding the real data.

C.2 Adaptation for monocular depth estimation

Translated datasets were also used for training unsupervised monocular depth estimation network (the same model as the one used for depth preservation) introduced by Godard et al. [6]. We find that even though our method improves the results over the baseline model, the results are worse than training on the source domain dataset. We believe that this is the case because the task network relies on consistency between left and right images by using a reconstruction loss on the reprojected images. In our case, the left and right images are translated independently, therefore any inconsistency between them is likely to affect the training. In Table 4 we show results from our observations. The models are evaluated on Kitti dataset as in Godard et al. [6]. Because of the different focal length and camera baseline in the dataset used in evaluation, we use a metric that we refer to as "scale aligned" absolute relative error. The estimated depth maps are aligned with the ground-truth by multiplying them by a single constant before calculating the absolute relative error. Such modified metric is more informative, because it is more independent of the scale of the disparities that depend on the camera parameters.

D Labels mapping

Even though all datasets we use correspond to driving scenes, semantic classes that they define are not the same. To address this problem we map the source labels as described in Table 5. Some classes in SYNTHIA do not have correspondances in Cityscapes and/or Microsoft COCO (denoted as NULL), therefore the ground-truth preservation loss (2) ignores such regions or we mask the gradients from those parts during back-propagation in the case of instance segmentation.

Table 4: Adaptation for monocular depth estimation. Evaluation on Kitti dataset. S = Semantic Segmentation, D = Depth.

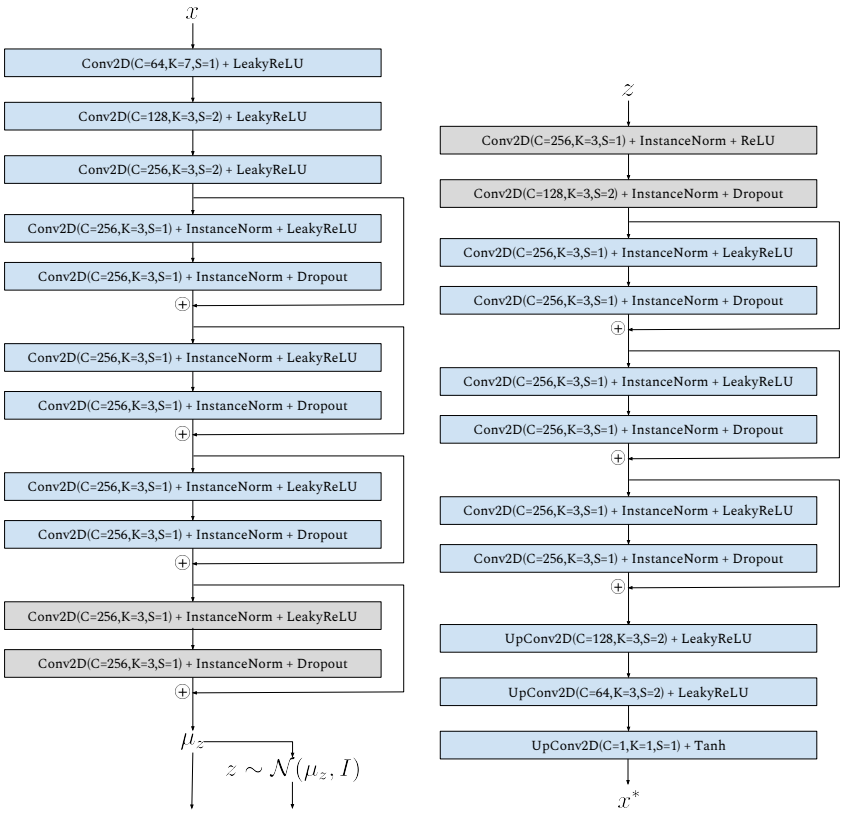
Dataset	"scale aligned" abs. rel. error
Source domain (SYNTHIA-Seq)	**0.2801**
GANtruth (S)	0.3401
GANtruth (D)	0.2849
GANtruth (S+D)	0.3051
UNIT	0.3417
UNIT + GANtruth (S)	0.3170
UNIT + GANtruth (D)	0.2862
UNIT + GANtruth (S+D)	0.2804

Table 5: Mapping of semantic classes from SYNTHIA to Cityscapes

(from) SYNTHIA		(to) Cityscapes		(to) Microsoft-COCO	
Class ID	Class	Class ID	Class	Class ID	Class
0	void	-	NULL	-	NULL
1	sky	10	sky	-	NULL
3	road	0	road	-	NULL
4	sidewalk	1	sidewalk	-	NULL
5	fence	4	fence	-	NULL
6	vegetation	8	vegetation	-	NULL
7	pole	5	pole	-	NULL
8	car	13	car	3	car
9	traffic sign	7	traffic sign	13	traffic sign
10	pedestrian	11	person	1	person
11	bicycle	18	bicycle	2	bicycle
12	lane-marking	0	road	-	NULL
13	reserved	-	NULL	-	NULL
14	reserved	-	NULL	-	NULL
15	traffic light	6	traffic light	10	traffic light

E Model architecture

The model architecture used in our experiments corresponds to the one proposed by Liu et al. [17] to make the comparisons fair. The details of the encoders and decoders are present in Figure 4. All of our discriminators are multi-scale. Each of them working on three different resolutions and have the architecture as described in Figure 5.



(a) Encoders – E_S, E_T (b) Decoders (generators) – G_S, G_T

Figure 4: Diagram of encoders and decoders of the models, modules marked as gray have the weights shared between domains. C, K and S correspond to the number of channels, kernel size and stride respectively

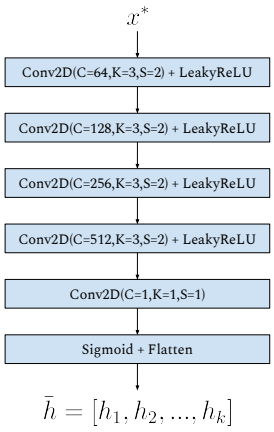


Figure 5: Detailed diagram of each discriminator – the module does not have any fully connected layers. The output $\bar{h}$ represents a probability on the input image x^* being true or fake, the number of output elements h_i depends on the input resolution. Each h_i has a receptive field of a different patch of the image. We use a multiscale discriminators – each of them works on the image of different resolution, but the general architecture of the discriminators is the same